

西方蜜蜂表皮蛋白基因 *apd-like* 转录起始位点的定位及 cDNA 序列的分析

孙亮先^{1,*}, 黄周英¹, 郑华军², 游燕琳¹

(1. 泉州师范学院, 分子生物学与药物化学福建省高等学校重点实验室, 福建泉州 362000;

2. 国家人类基因组南方研究中心, 上海 201203)

摘要: Apidermin 蛋白家族是根据蜜蜂表皮蛋白 apidermin 1–3 (APD 1–3) 而命名的一个新型的昆虫结构性表皮蛋白家族。为了鉴定西方蜜蜂 *Apis mellifera* 基因组序列上毗邻基因簇 *apd* 1–3 的一个预测基因座 LOC727145 是否为一个新的 *apd* 基因, 本研究在用 5' LongSAGE 标签定位该基因的转录起始位点 (TSS) 的基础上, 利用其中的 3 条 5' LongSAGE 标签序列作为上游引物, 通过 RT-PCR 方法克隆了该基因的 cDNA 序列 (GenBank 登录号: GU358197, GU358199, GU358198)。生物信息学分析发现, 基因座 LOC727145 含有 2 个外显子和 1 个“GT-AG”型内含子, 其 cDNA 序列富含 GC (70%), 可编码一条长 152 aa 残基的高度疏水性多肽。此多肽序列的氨基酸组成与蜜蜂 APD 1–3 表皮蛋白类似, 富含 Ala, Gly, Pro, Leu 和 Val 5 种氨基酸 (占 77%), 其中 Ala 残基含量最高 (29%)。该多肽序列与蜜蜂 APD-1 表皮蛋白序列的相似性为 50%, 且其 N 末端的预测信号肽序列与 APD 蛋白的信号肽序列类似。5' LongSAGE 标签的基因组定位结果显示, 基因座 LOC727145 在雄蜂头部中表达丰度很高, RNA Pol II 可从 6 个不同的 TSS 上以不同效率起始转录, 其中由一个优势 TSS 上起始了 90% 的转录。本研究为 apidermin 表皮蛋白家族增添了一个新成员, 命名为 apidermin-like (*apd-like*)。

关键词: 西方蜜蜂; 5' LongSAGE; 表皮蛋白; *apd-like* 基因; 转录起始位点 (TSS)

中图分类号: Q966 文献标识码: A 文章编号: 0454-6296 (2011) 02-0224-08

TSS identification and cDNA cloning of a cuticular protein gene *apd-like* from the Western honeybee (*Apis mellifera*)

SUN Liang-Xian^{1,*}, HUANG Zhou-Ying¹, ZHENG Hua-Jun², YOU Yan-Lin¹ (1. Molecular Biology and Pharmacology Key Laboratory of Fujian Advanced Education, Quanzhou Normal University, Quanzhou, Fujian 362000, China; 2. Chinese National Human Genome Center at Shanghai, Shanghai 201203, China)

Abstract: The apidermin (APD) family, named after three cuticular proteins (apidermin 1–3, APD 1–3) in the Western honeybee, *Apis mellifera*, is a novel insect structural cuticular protein family. A hypothetical gene locus (LOC727145) was found to be located at the upstream of the reported *apd* gene cluster in the honeybee genomic sequence. In order to characterize this gene, we mapped its transcription start sites (TSSs) by a group of 5' LongSAGE tag sequences, and subsequently cloned its cDNA sequences through RT-PCR by using three different 5' LongSAGE tag sequences as the up-stream primers. The cloned cDNA sequences were submitted to GenBank under the accession numbers GU358197, GU358198 and GU358199. Bioinformatics analysis revealed that the genomic DNA sequence LOC727145 contains two exons and one intron, where the boundary conforms to the GU/AG rule. The cDNA sequence is extremely GC-rich (70%), and the deduced protein is highly hydrophobic. Five amino acids (Ala, Gly, Pro, Leu and Val) account for 77% of the deduced polypeptide sequence, and Ala is the most abundant residue (29%). These characters are comparative to APD 1–3. Moreover, the deduced polypeptide sequence has 50% identity with APD-1 protein, and both possess a similar predicted hydrophobic signal peptide at the N-terminal. These results suggest that LOC727145 encodes a novel APD protein. The mapping result of 5' LongSAGE tag sequences on genome sequence showed that LOC727145 was highly expressed in drone head, and RNA Pol II initiated transcription of this gene at six alternative TSSs with different efficiency,

基金项目: 国家自然科学基金项目 (30300183, 31072095); 福建省教育厅省属高校项目 (2008F5051)

作者简介: 孙亮先, 1970 年生, 山东胶州人, 博士, 副教授, 主要从事蜜蜂功能基因组学研究

* 通讯作者 Corresponding author, E-mail: lxsun@sohu.com

收稿日期 Received: 2010-06-21; 接受日期 Accepted: 2011-12-20

and 90% of the mRNA molecules came from a dominant TSS. This research identified a novel apidermin protein, named apidermin-like (*apd-like*).

Key words: *Apis mellifera*; 5'LongSAGE; cuticular protein; *apd-like* gene; transcription start site (TSS)

表皮蛋白是昆虫表皮的重要组分。表皮蛋白有的会与几丁质发生交联,有的会被儿茶酚衍生物修饰,有的本身具有很强的疏水性,因此有相当数量的表皮蛋白很难提取。如甲虫 *Agrianome spinnicollis* 幼虫中无法提取的表皮蛋白占 56%, 只能通过水解除去;而蜕皮后尚未硬化的蝗虫表皮中不能提取的蛋白占总蛋白的 10% (Andersen *et al.*, 1995)。对于无法提取的昆虫表皮蛋白,无法在蛋白水平上直接分析其结构和性质。

近年来,越来越多的昆虫表皮蛋白的序列被报道。Andersen 等(1995)综述了 38 条昆虫表皮蛋白序列的特征。至 2005 年,已报道的表皮蛋白已达 139 条,其中 74 条是真正的表皮蛋白,而另外 65 条则是由 cDNA, ESTs (expressed sequence tags) 和基因组 DNA 片段序列推导而来,然后根据与已知表皮蛋白序列的相似性而将其产物蛋白进行归类 (Willis *et al.*, 2005)。在黑腹果蝇 *Drosophila melanogaster*、冈比亚按蚊 *Anopheles gambiae*、西方蜜蜂 *Apis mellifera*、家蚕 *Bombyx mori*、金小蜂 *Nasonia* 等昆虫的全基因组序列发表之后,基于与已知表皮蛋白序列的相似性,研究人员从基因组序列中鉴别到数以百计的表皮蛋白基因(冈比亚按蚊, 240 个;家蚕, 221 个;黑腹果蝇, 155 个;西方蜜蜂, 47 个;金小蜂, 76 个) (Willis, 2010)。He 等(2007)使用 LC/MS/MS 技术分析了冈比亚按蚊的表皮蛋白组,证实了 90% 的根据基因组序列注释的表皮蛋白基因的产物蛋白可在表皮内检测到,说明从基因组序列中所注释的昆虫表皮蛋白基因具有较高的可信度。最近,Willis (2010)综述了目前所有昆虫表皮蛋白的序列特征,并将其划分为 12 个家族。从原理上看,只要获得了基因的基因组 DNA 序列或其 cDNA 序列,就可以使用生物信息学方法分析其产物蛋白的是否含有某个表皮蛋白家族的共有基序(motif),从而判断该蛋白是否为表皮蛋白。这种从核酸序列着手鉴定表皮蛋白的策略对于分析那些难以纯化的表皮蛋白尤为有用。

蜜蜂是一种重要的农业有益昆虫,也是研究社会行为学和神经生物学的重要模式生物。在对西方蜜蜂 *Apis mellifera* 全基因组序列进行功能注释时,蜜蜂基因组测序联盟预测到 28 个编码 R&R 表皮蛋

白的基因,其数量仅相当于果蝇或蚊子该类基因数量的三分之一(The Honeybee Genome Sequencing Consortium, 2006)。之后,Kucharski 等(2007)报道了 3 个被命名为 *apidermin* (*apd 1-3*) 的表皮蛋白基因。这 3 个基因串联排列在第 4 号染色体的一段富含 GC 的区域内,其预测产物蛋白高度疏水,且不含 R&R 保守结构域。从氨基酸组成上看,这 3 条 APD 多肽序列中 Ala, Gly, Leu, Pro 和 Val 等 5 种氨基酸的含量为 74% ~ 86%,其中 Ala 残基含量均超过 30%。*apd 1-3* 基因在工蜂中具有时空特异性表达的特点,其中 *apd-1* 在老熟蛹和新羽化蜂的外骨骼上皮中表达,*apd-2* 在气管、食道、胃和胚胎内表达,而在白蛹期的外表皮和眼表皮内则检测到 *apd-3* 的表达。最近,Zheng 和 Sun 使用西方蜜蜂成年雄蜂头部 5' LongSAGE 文库中的 8 472 条不同的标签序列在蜜蜂全基因组序列上注释了 3 966 个预测基因的转录起始位点(TSS)(未发表数据)。我们的 5'LongSAGE 数据显示,*apd-2* 基因在雄蜂头部的表达丰度非常高,共被检测到 416 次;*apd-3* 基因的表达丰度较低,只检测到 1 次,而未检测到 *apd-1* 的转录(未发表数据)。有趣的是,我们发现毗邻蜜蜂 *apd* 基因簇的一个预测基因座 LOC727145 的转录水平与 *apd-2* 基因相似(共检测到 429 次),有 6 条不同的 5'LongSAGE 标签序列被定位在该预测基因的 TSS 周围。由于简单基因家族的成员往往在基因组上串联排列,因此 LOC727145 很可能是一个新的 *apd* 基因。为此,本研究利用定位在 LOC727145 的 3 条 5'LongSAGE 标签序列作为 PCR 上游引物,通过 RT-PCR 克隆了该基因的 cDNA 序列,并进而分析了该基因的结构及其产物蛋白的结构特点。我们的分析结果显示 LOC727145 编码一个新的 APD 表皮蛋白(命名为 apidermin-like, *apd-like*)据我们所知,*apd-like* 是继 *apd 1-3* 之后的第 4 个有转录数据支持的 APD 表皮蛋白基因。

1 材料与方法

1.1 蜜蜂样品及总 RNA 的提取

意大利蜜蜂 *Apis mellifera ligustica* 蜂群饲养在泉州师范学院模式生物研究中心的实验蜂场。从巢

脾上随机捉雄蜂,置于 -80°C 冰箱内备用,用于提取总 RNA。切取头部,液氮研磨后按 $50\sim 100\text{ mg}$ 的样品加入 1 mL TRIzol 试剂(Gibco BRL 公司),参照 Trizol 试剂盒说明书提取总 RNA。取 $1\text{ }\mu\text{L}$ 总 RNA 溶液于 1% 琼脂糖凝胶电泳检测。

1.2 反转录合成第一链 cDNA

取一个 RNase-free 的离心管(AMBION 公司),加入 1 mg 总 RNA, $17.4\text{ }\mu\text{L}$ RT-PCR 级水(Ambion 公司), $0.6\text{ }\mu\text{L}$ Oligo-dT 引物($0.5\text{ }\mu\text{g}/\mu\text{L}$, Invitrogen 公司),混匀后 65°C 水浴 10 min ,然后降温至 37°C ,停留 1 min 后升温至 42°C 。另取离心管,加入 StrataScript buffer $5\text{ }\mu\text{L}$, dNTP(10 mmol/L dNTP, Invitrogen 公司) $5\text{ }\mu\text{L}$, StrataScript Reverse Transcriptase ($50\text{ U}/\mu\text{L}$, Stratagen 公司) $1\text{ }\mu\text{L}$, StrataScript Reverse Transcriptase Inhibitor($40\text{ U}/\mu\text{L}$, Stratagen 公司) $1\text{ }\mu\text{L}$ 。混匀后加入第一离心管中, 42°C 1 h 后, 90°C 5 min 灭活逆转录酶。

1.3 PCR 反应及 PCR 产物的克隆

1.3.1 引物设计:以西方蜜蜂雄蜂 5'LongSAGE 文库(GenBank 登录号: GSE19555)中的标签 5'-AGCACATCCACCCGAGCCA-3'(GenBank 登录号: AMDH2220)、5'-ACAGCACATCCACCCGAGC-3'(GenBank 登录号: AMDH1229)和 5'-ACATCCACCCGAGCCACCG-3'(GenBank 登录号: AMDH1308)作为 PCR 上游引物序列;下游引物根据标签 AMDH2220 所定位的预测基因(LOC727145)的最后一个外显子序列设计,序列为 5'-CAGCTGCTATGCTACCCGAG-3'(图 1: B)。引物委托上海生工合成。

1.3.2 PCR 反应与产物检测:PCR 反应体系为: LA Taq (TaKaRa 公司) $0.25\text{ }\mu\text{L}$, $2\times\text{ GC Buffer II}$ $20\text{ }\mu\text{L}$, 2.5 mmol/L dNTP(TaKaRa 公司) $4\text{ }\mu\text{L}$, 模板(反转录产物) $6.25\text{ }\mu\text{L}$, 上游和引物($2\text{ pg}/\mu\text{L}$)($2\text{ pg}/\mu\text{L}$)各 $1\text{ }\mu\text{L}$ 。反应条件为: 95°C 变性 3 min ; 95°C 30 s , 57°C 30 s , 72°C 2 min ; 35 cycles ; 72°C 保温 10 min 。取 $5\text{ }\mu\text{L}$ PCR 产物于 1.5% 琼脂糖凝胶电泳检测,在一个泳道中加 1 kb Plus DNA Ladder (Tiangen 公司)。

1.3.3 PCR 产物的纯化与克隆:将 PCR 产物在 1.5% 琼脂糖凝胶电泳后,于紫外灯下切取含目的 DNA 条带的琼脂糖胶块,以 QIAquick Gel Extraction Kit(Qiagen 公司)回收纯化目的 DNA 片段。使用 pMD18-T 试剂盒(TaKaRa 公司)克隆 PCR 产物,具体操作步骤参照试剂盒说明书。反应体系为: Vector $0.2\text{ }\mu\text{L}$, Solution I $3\text{ }\mu\text{L}$, 回收 PCR 产物

$4.8\text{ }\mu\text{L}$ 。 16°C 水浴 1 h 后 70°C 灭活 10 min 。取 $2\text{ }\mu\text{L}$ 连接液转化大肠杆菌 *Escherichia coli* DH10B 菌株的感受态细胞, 37°C 220 rpm 培养 1 h 后,在 1 mL 细菌培养液中加入 $20\text{ }\mu\text{L}$ IPTG(100 mg/mL),混匀后涂布 LB 平板(含有 amp 抗生素, X-gal), 37°C 恒温培养过夜。

1.3.4 重组子的菌液 PCR 鉴定:在 PCR 管中加入 $20\text{ }\mu\text{L}$ 液体 LB 培养基,接种白色单克隆菌落, 37°C 培养 60 min ,取菌液作为 PCR 扩增的模板。所用引物序列为 pMD18-T 载体上的测序通用引物 P2(5'-GTAAAACGACGGCCAGT-3')和 P4(5'-AGCGGATAACAATTTTCACACAGGA-3')。使用 Golden Easy PCR System(Tiagen)进行 PCR 反应。反应体系为: $2\times\text{ Reaction Mix}$ $5\text{ }\mu\text{L}$, 菌液(模板) $2\text{ }\mu\text{L}$, 引物各 $1\text{ }\mu\text{L}$, ddH₂O $1\text{ }\mu\text{L}$, Taq DNA 聚合酶 $0.2\text{ }\mu\text{L}$ ($2.5\text{ U}/\mu\text{L}$)。PCR 反应条件为: 95°C 变性 5 min ; 95°C 30 s , 57°C 40 s , 72°C 2 min , 30 轮循环 ; 72°C 保温 10 min 。取 $5\text{ }\mu\text{L}$ PCR 产物于 1% 琼脂糖凝胶电泳检测。

1.3.5 重组克隆的质粒抽提和测序:取菌液 PCR 鉴定为阳性的菌液,在含有 Amp 抗生素的 LB 固体培养基划线, 37°C 培养过夜。挑取白色单克隆菌落,接种 5 mL 液体 LB 培养基(Amp $100\text{ }\mu\text{g/mL}$)于 37°C , 300 rpm 培养过夜。以质粒小提试剂盒(Tiagen 公司)抽提质粒。以 pMD18-T 载体上的测序通用引物 P₂ 和 P₄ 对所克隆的 PCR 产物进行双向测序。

1.4 生物信息学分析

1.4.1 基因座 LOC727145 的基因组 DNA 序列及 5'LongSAGE 标签:用于定位 LOC727145 转录起始位点的 5'LongSAGE 标签来自于 NCBI 的雄蜂 5'LongSAGE 文库(Series record GSE19555),标签序列为: 5'-GACAGCACATCCACCCGAG-3', 5'-ACAGCACATCCACCCGAGC3', 5'-AGCACATCCACCCGAGCCA-3', 5'-GCACATCCACCCGAGCCAC-3', 5'-ACATCCACCCGAGCCACCG-3', 5'-ATCCACCCGAGCCACCGAG-3'。LOC727145 所在的 Contig 序列(Contig1775)及该基因的预测的 cDNA 序列(GenBank 登录号: XM_001122858)和多肽序列(GenBank 登录号: XP_001122858.1)从 NCBI 网站下载。

1.4.2 启动子的分析:在用 5'LongSAGE 标签在 Contig1775 中定位 LOC727145 的 TSS 后,用 DNASTAR 软件截取 TSS 上游 300 bp 至下游 100 bp 的序列,使用 TFSEARCH server 网站(<http://>

www. cbr. jp/research/db/TFSEACRCH) 在线分析启动子中可能含有的调控元件。

1.4.3 cDNA 序列的分析：用 phred/Phrap/consed 软件对 RT-PCR 产物的测序结果进行拼装，得到 cDNA 的序列。使用 DNASTAR 软件分析该 cDNA 序列的碱基组成、所编码的多肽及多肽的氨基酸组成。信号肽的预测在 SignalP server 网站上进行 (<http://www.cbs.dtu.sk/services/SignalP>)。多肽的多重比较使用 MEGA 4.0 软件进行。在 NCBI 网站上使用 BLAST 软件用 cDNA 序列对蜜蜂基因组序列 (Amel_4.0) 注释 LOC727145 的内含子\外显子结构。

2 结果与分析

2.1 西方蜜蜂基因组序列上基因座 LOC727145 的 cDNA 序列分析

为了验证西方蜜蜂基因组序列上定位于预测基因座 LOC727145 的 5'LongSAGE 标签确实来自于该基因的 mRNA 分子，本研究分别选取拷贝数为 13, 388 和 17 的 3 条标签序列作为 PCR 的上游引

物，下游引物序列根据该基因的最后一个外显子设计 (图 1: B)，以雄蜂头部的总 RNA 为模板，通过 RT-PCR 获得 3 条长约 500 bp 的 DNA 片段。图 2 (A) 为拷贝数为 388 的标签序列作为上游引物的扩增产物的琼脂糖凝胶电泳结果，图 2(B) 为此 cDNA 片段克隆后的菌落 PCR 产物的电泳结果。抽提质粒进行测序分析，得到了 3 条长分别为 487 bp, 485 bp 和 482 bp 的 DNA 序列 (GenBank 登录号分别为：GU358197, GU358198, GU358199)。这 3 条 cDNA 序列的两个末端均为引物序列，其序列除在 5' 端长度不同外，其他碱基序列完全相同。此 cDNA 序列的一个显著特征是富含 GC (70.3%)。据 Kucharski 等 (2007) 报道，西方蜜蜂表皮蛋白基因 *apd 1-3* 的编码区的 GC 含量均高于 50%，其中 *apd-1* 基因编码区的 GC 含量超过 70%。在基因组 DNA 水平上，*apd* 基因簇所在区域的 GC 含量为 38.5%，而基因座 LOC727145 的基因组 DNA 的 GC 含量为 55.6%；而蜜蜂全基因组序列中的 GC 含量为 33%，基因所在区域的平均 GC 含量仅为 29% (The Honeybee Genome Sequencing Consortium, 2006)。

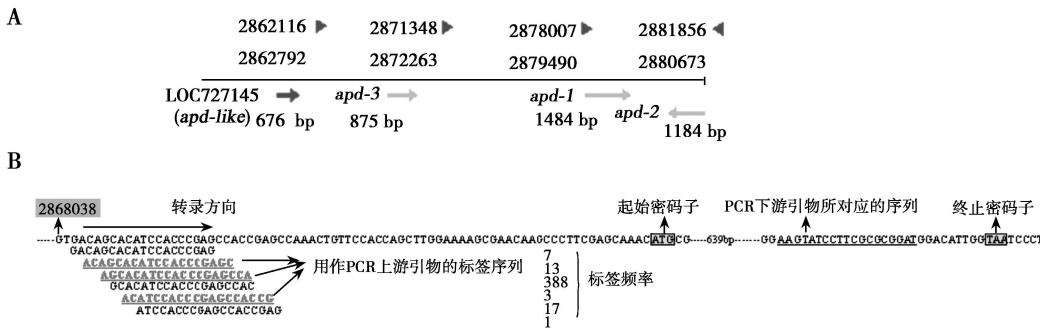


图 1 5'LongSAGE 标签在西方蜜蜂基因组序列上所定位的基因座 LOC727145 的转录起始位点

Fig. 1 The transcription start sites (TSSs) of the hypothetical locus LOC727145 of *apd-like* gene from *Apis mellifera* mapped by 5' LongSAGE tags

A: LOC727145 在 Contig1775 中位于表皮蛋白 *apd 1-3* 的上游，长箭头表示转录方向，短箭头表示编码序列的起始位置。The genomic context of hypothetical gene locus LOC727145. It is located at the upstream of three apidermin genes (*apd 1-3*). The long arrow indicates the direction of transcription and the short arrows indicate the initiation site of start codon. B: 定位在 LOC727145 转录起始位点周围的 5'LongSAGE 标签序列。标签来自于西方蜜蜂成年雄蜂头部的 5'LongSAGE 文库，标签数量代表 mRNA 由该位点起始转录的频率，编码序列的起始密码子 ATG 和终止密码子 TAA 用黑体标志，用作 RT-PCR 引物序列的 5'LongSAGE 标签用下划线标志。The tag hits around transcriptional start site of locus LOC727145. The tags are derived from a 5'LongSAGE library of adult drone head. The number of tags indicates the initiation efficiency. The start codon and stop codon are in bold. The sequences used for designing PCR primers are underlined.

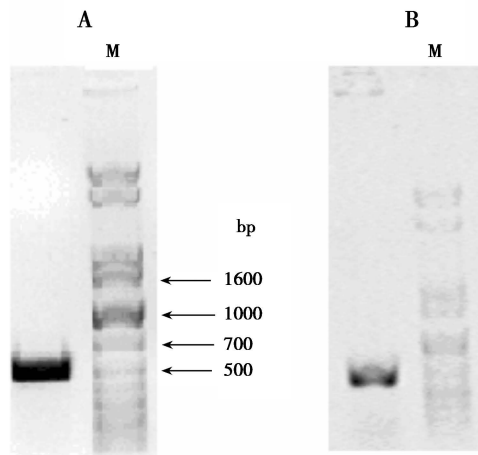


图2 西方蜜蜂基因组上基因座 LOC727145 的 RT-PCR 产物(A)与其重组子菌落 PCR(B)琼脂糖凝胶电泳结果
Fig. 2 Agarose gel electrophoresis result of RT-PCR product of locus LOC727145 on *Apis mellifera* genome (A) and the corresponding recombinant clone (B)
M: DNA ladder.

表 1 比较了用 cDNA 序列 (GenBank 登录号:

表 1 用 cDNA 序列和软件预测 mRNA 所注释的西方蜜蜂基因组序列上基因座 LOC727145 的外显子
Table 1 Exons of locus LOC727145 on genomic sequences of *Apis mellifera* that were annotated with 5'LongSAGE tag-generated cDNA and *in silico* predicted mRNA

	GenBank 登录号 GenBank accession no.	cDNA 长度(bp) Length of cDNA	所注释的外显子在基因组中的位置 Location of the exons annotated on genomic DNA	
cDNA	GU358199	485	1 - 387	386 - 485
基因组 DNA Genomic DNA	Contig1775		1 660 387 - 1 660 773	1 660 990 - 1 661 089
预测的 mRNA Predicted mRNA	XM_001122858	459	1 - 314	313 - 459
基因组 DNA Genomic DNA	Contig1775		1 660 460 - 1 660 773	1 660 990 - 1 661 136

2.2 西方蜜蜂基因组上基因座 LOC727145 转录起始位点的定位

在西方蜜蜂全基因组的功能注释中,预测基因座 LOC727145 位于连锁群 4.18,下游毗邻蜜蜂表皮蛋白基因簇 *apd 1-3* (图 1: A)。在向 GenBank 释放了该基因的 cDNA 序列后 (GenBank 登录号: GU358197, GU358199, GU358198; 10-MAR-2010), NCBI 在 4 月 10 日按照我们的注释将该基因座的功能注释更新为 *apidermin-like* (4-Apr-2010)。因为 5' LongSAGE 标签来自于全长 mRNA 分子的 5' 末端,因此 5' LongSAGE 文库中不同标签序列在基因组上的位置可代表相应活性基因的 TSS,而所检测到的各标签序列的拷贝数则可代表由该 TSS 起始转录的效率 (Wei *et al.*, 2004)。5' LongSAGE 标签序列的

GU358199) 与预测的参考 mRNA 序列 (GenBank 登录号: XM_001122858) 对基因座 LOC727145 所注释的外显子。二者所注释的基因结构基本相同,都是由 2 个外显子和 1 个长 218 bp 内含子组成 (表 1)。此内含子的边界序列为“5' GG ↓ GTG---TAG ↓ GA 3'”,属于典型的“GU-AG”型内含子。从碱基组成上看,LOC727145 的一个明显的特征是其内含子富含 AT (68%) 而外显子却富含 GC (70%)。与软件预测的两个外显子相比,用克隆到的 cDNA 序列所注释的第一个外显子在 5' 端长了 74 bp,而第二个外显子在 3' 端短 48 bp (表 1)。这是由于蜜蜂基因组测序联盟所预测的转录单元是由起始密码子开始至终止密码子结束 (The Honeybee Genome Sequencing Consortium, 2006),而本研究所用的 RT-PCR 上游引物序列在 5' 末端包含了转录起始位点,而下游引物则位于预测基因最后一个外显子的内部,因此所克隆的 cDNA 在 5' 端含有 mRNA 的 5' UTR 序列,但 3' 端尚不完整。

基因组定位结果显示,在雄蜂的头部中,*apd-like* 基因可从位于起始密码子上游的 76 bp 范围内的 6 个 TSS 上起始转录,并且所转录的第一个碱基都是嘌呤 (图 1: B)。RNA 聚合酶 II 从这 6 个 TSS 上以不同的效率起始转录,其中,由标签序列“AGCACATCCACCCGAGCCA” (388 个拷贝) 所定位的 TSS 为优势 TSS, *apd-like* 基因的 90.4% 的 mRNA 分子由该位点起始。此结果显示在雄蜂头部中 *apd-like* 基因可从 6 个选择性转录起始位点上 (alternative TSS) 起始转录,且其表达水平非常高。

2.3 西方蜜蜂 APD-like 蛋白序列的分析

基因座 LOC727145 的预测产物蛋白 (GenBank 登录号: XP_001122858.1) 含有 152 个氨基酸残基,其中疏水性氨基酸占 72%,说明此蛋白具有高度的

疏水性。从氨基酸组成上看, Ala, Gly, Pro, Leu 和 Val 5 种氨基酸占这条多肽序列的 77%, 其中 Ala 含量最高 (29%), Gly 为第二丰富的氨基酸 (16%)。而已报道的 3 条蜜蜂表皮蛋白 APD 1-3 也具有高度疏水性, 其中 Ala 含量均高于 30% (Kucharski *et al.*, 2007)。可见 LOC727145 所编码的蛋白在氨基酸组成上与 APD 1-3 非常相似。

LOC727145 产物蛋白的 N 末端有一段 18 AA 的预测信号肽序列, 该信号肽能形成 α -螺旋结构 (图 3), 暗示着此蛋白是一种分泌型蛋白。APD-like 多肽序列中存在低重复数的短重复序列。如 AALAGPI/A 重复出现了 3 次, 其疏水性氨基酸会形成 α -螺旋二级结构, 而在短重复序列中的 Pro 残基处则可能形成 α -转角, 从而改变肽链走向。用 BLAST 软件在 NCBI 网站上在线搜索数据库时, 没

有在 APD-like 序列中发现保守的结构域; 但在 Swissprothttp://expasy.org/tools/ 网站上用 ScanProsite 软件在线搜索时发现此多肽序列的 5 ~ 148 位氨基酸序列是一个保守的富含丙氨酸区 (Ala-rich region profile)。本研究没有在 APD-like 中发现昆虫表皮蛋白的特征性基序, 但蛋白序列比对结果显示 (图 4), APD-like 蛋白与蜜蜂表皮蛋白 APD-1, APD-3 和 APD-2 的同源性分别为 50%, 42% 和 32%。由于这 4 种蛋白具有序列类似, N 末端都有一条氨基酸组成类似的信号肽, 多肽内均富含疏水性氨基酸等相似特征, 而且这 4 个蛋白的编码基因串联排列成一个基因簇, 属于同一蛋白家族, 因此我们将之命名为 apidermin-like (APD-like)。

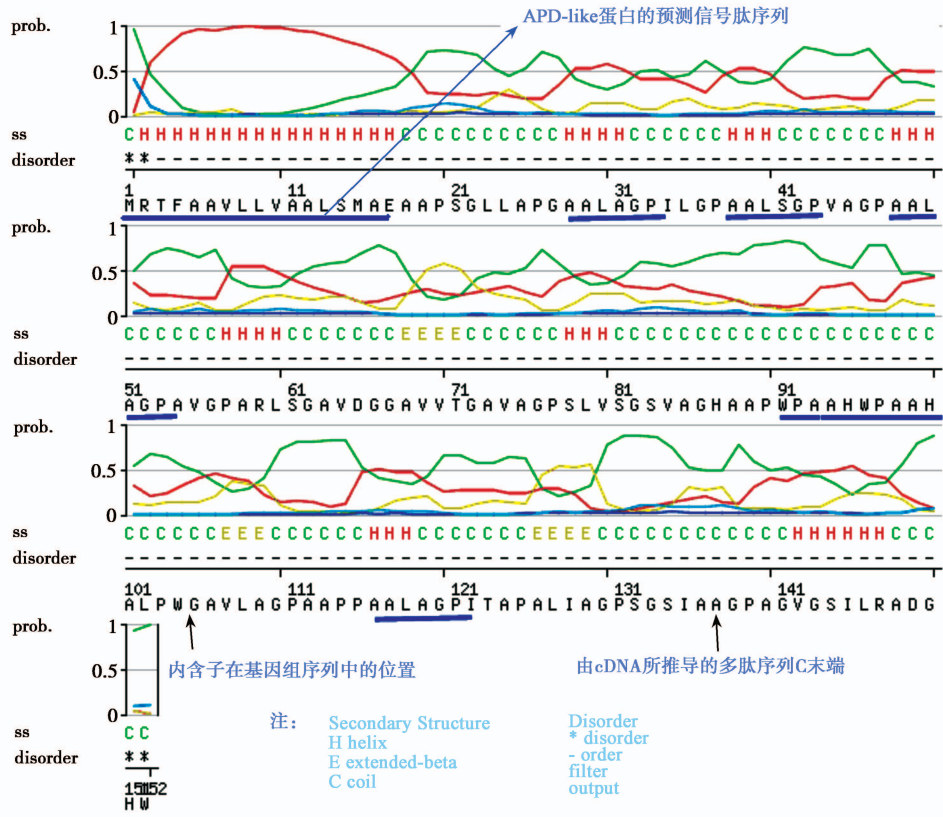


图 3 西方蜜蜂 APD-like 蛋白的序列和预测的二级结构

Fig. 3 The polypeptide sequence of APD-like protein from *Apis mellifera* and its predicted secondary structure
预测的信号肽序列用带箭头的下划线表示, 多肽序列中的短重复序列用下划线表示, 由 cDNA 所推导的多肽序列 C 末端和内含子在基因组序列中的位置用箭头表示。The predicted signal peptide is underlined. The short sequence repeats are underlined. The arrows indicate the position of intron in the genomic sequence and the C-terminal of polypeptide sequence that is deduced from RT-PCR amplified cDNA.



图4 西方蜜蜂 APD-like 蛋白与已报道的蜜蜂表皮蛋白 APD 1-3 的多肽序列比较

Fig. 4 Alignment of APD-like protein from *Apis mellifera* with the reported honeybee cuticular proteins APD 1-3

2.4 西方蜜蜂 *apd-like* 基因启动子区域中的顺式作用元件

转录起始位点(TSS)是基因组序列中各转录单元的5'边界。对于真核生物的蛋白质编码基因而言,启动子元件一般位于TSS上游300 bp至下游50 bp之内。只有少数基因在TSS上游几kb的位置也存在调控元件,而超出TSS下游100 bp的范围内很少存在启动子元件。理论上只要能在DNA序列上准确地定位TSS,就可以准确地分析该基因的启

动子序列。因此,在定位了 *apd-like* 基因的6个选择性TSS之后,本研究对TSS附近的-300 bp ~ +100 bp区域内的顺式作用元件进行了在线分析。分析结果显示(图5),在第一个TSS上游-20 ~ -30区存在一个典型的TATA框,在-59 ~ -65和-87 ~ -93区为2个GC框;此外,在下游+34 ~ +39区和下游-107 ~ -117区、-272 ~ -286区有3个热休克转录因子(HSF)结合位点。

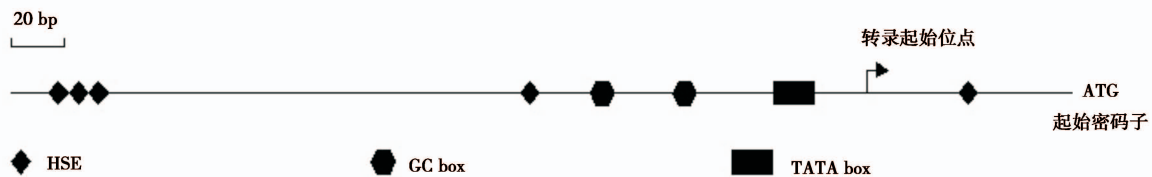


图5 西方蜜蜂 *apd-like* 基因5'侧翼序列上预测的顺式作用元件的分布

Fig. 5 Scheme of putative *cis*-acting elements mapped on the 5'-flanking sequence of *apd-like* gene from *Apis mellifera*

箭头表示用5' LongSAGE 标签定位的转录起始位点,ATG表示起始密码子。The transcription start sites are mapped by 5' LongSAGE tags. ATG is the start codon.

3 讨论

5' LongSAGE 技术是利用CAP-Trapper原理获得含有5'帽子的全长cDNA,然后通过一系列的加接头、酶切、连接、克隆和高通量测序等操作,最终得到转录组中各种mRNA分子的5'末端的长19 nt的标签序列。将这些标签序列定位在全基因组序列上,就可以高通量地获得转录单元的TSS信息,而标签序列被检测到的次数则可代表转录组中各mRNA分子的丰度(Wei *et al.*, 2004)。本研究发现西方蜜蜂雄蜂5'LongSAGE文库中的一组标签序列可定位在蜜蜂基因组序列的一个预测基因座(LOC727145)的编码序列上游,从而证实了基因组中确实存在这个基因座。引人注目的是,在雄蜂头部转录组中,此基因座的转录水平仅次于位于它的上游的一个预测基因座(LOC409716)。而且,基因

座LOC727145在雄蜂头部可从6个TSS上以不同的效率起始转录。在查看LOC727145在基因组的位置时,我们发现该基因位于第4号染色体上,其上游毗邻表皮蛋白基因簇*apd 1-3*。在对其序列进行分析时,我们发现基因座LOC727145的编码序列富含GC(70%),所编码的蛋白与蜜蜂APD-1有较高的同源性为50%,且此多肽序列具有高度疏水、富含Ala、N末端有疏水性的信号肽序列等表皮蛋白基因*apd 1-3*的特征。因为真核生物简单基因家族中的基因一般以串联方式前后相连,因此根据以上证据,可以我们认为基因座LOC727145与蜜蜂表皮蛋白基因*apd 1-3*属于同一基因家族。

由于5'LongSAGE标签来自于mRNA的5'末端的序列,所以在理论上可作为克隆全长cDNA的5'端PCR引物。为了验证基因座LOC727145的预测结构是否正确,本研究尝试用定位在该基因的5' LongSAGE 标签序列作为上游引物,而下游引物根

据最后一个外显子序列设计,通过 RT-PCR 克隆 cDNA,结果得到了 3 条具有完整 5'末端的 cDNA 序列。此结果不仅验证了 5'LongSAGE 标签序列确实是来自于 mRNA 分子,而且也证明使用 5' LongSAGE 标签序列作为 RT-PCR 引物克隆全长 cDNA 的 5'端序列策略的可行性。相对于文库杂交筛选、RACE 等方法,该 cDNA 克隆策略更为经济快捷。

参 考 文 献 (References)

- Andersen SO, Hojru PP, Roepstorff P, 1995. Insect cuticular proteins. *Insect Biochem. Mol. Biol.*, 25(2): 153 – 176.
- He N, Botelho JMC, McNall RJ, Belozarov V, Dunn WA, Mize T, Orlando R, Willis JH, 2007. Proteomic analysis of cast cuticles from *Anopheles gambiae* by tandem mass spectrometry. *Insect Biochem. Mol. Biol.*, 37: 135 – 146.
- Kucharski R, Maleszka J, Maleszka R, 2007. Novel cuticular proteins revealed by the honey bee genome. *Insect Biochem. Mol. Biol.*, 37: 128 – 134.
- The Honeybee Genome Sequencing Consortium, 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature*, 443: 931 – 949.
- Wei CL, Ng P, Chiu KP, Chiu KP, Wong CH, Ang CC, Lipovich L, Liu ET, Ruan Y, 2004. 5' Long serial analysis of gene expression (LongSAGE) and 3' LongSAGE for transcriptome characterization and genome annotation. *Proc. Natl. Acad. Sci. USA*, 101(32): 11701 – 11706.
- Willis JH, 2010. Structural cuticular proteins from arthropods: annotation, nomenclature, and sequence characteristics in the genomics era. *Insect Biochem. Mol. Biol.*, 40(3): 189 – 204.
- Willis JH, Iconomidou VA, Smith RF, Hamodrakas SJ, 2005. Cuticular proteins. In: Gilbert LI, Latrou K, Gill SS eds. *Comprehensive Molecular Insect Science*. Vol. 4. Oxford, Elsevier. 79 – 109.

(责任编辑: 赵利辉)